University of Edinburgh

School of Mathematics

Bayesian Data Analysis, 2022/2023, Semester 2

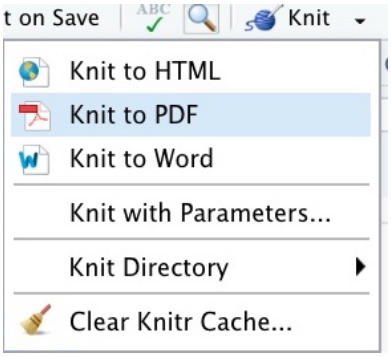
Assignment 2

IMPORTANT INFORMATION ABOUT THE ASSIGNMENT

In this paragraph, we summarize the essential information about this assignment. The format and rules for this assignment are different from your other courses, so please pay attention.

1) Deadline: The deadline for submitting your solutions to this assignment is the 17 April 12:00 noon Edinburgh time.

2) Format: You will need to submit your work as 2 components: a PDF report, and your R Markdown (.Rmd) notebook. There will be two separate submission systems on Learn: Gradescope for the report in PDF format, and a Learn assignment for the code in Rmd format. You need to write your solutions into this R Markdown notebook (code in R chunks and explanations in Markdown chunks), and then select Knit/Knit to PDF in RStudio to create a PDF report.



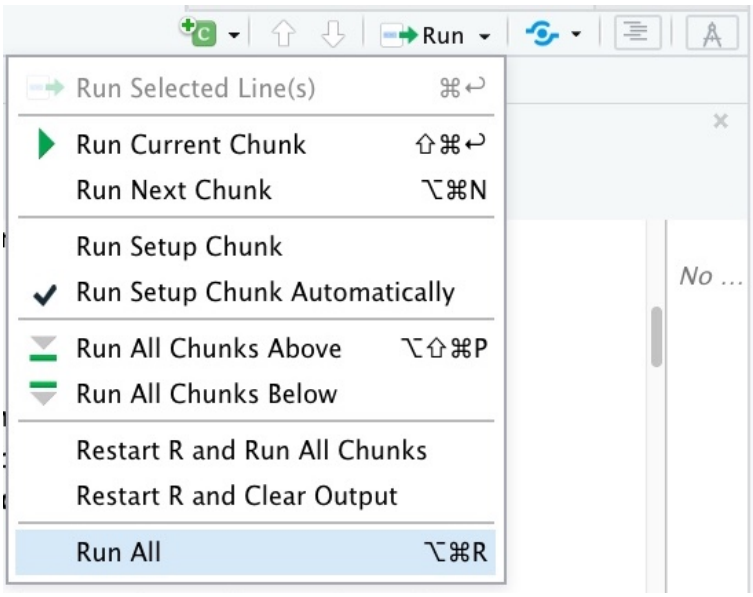
The compiled PDF needs to contain everything in this notebook, with your code sections clearly visible (not hidden), and the output of your code included. Reports without the code displayed in the PDF, or without the output of your code included in the PDF will be marked as 0, with the only feedback "Report did not meet submission requirements".

You need to upload this PDF in Gradescope submission system, and your Rmd file in the Learn assignment submission system. You will be required to tag every sub question on Gradescope.

Some key points that are different from other courses:

a) Your report needs to contain written explanation for each question that you solve, and some numbers or plots showing your results. Solutions without written explanation that clearly demonstrates that you understand what you are doing will be marked as 0 irrespectively whether the numerics are correct or not.

b) Your code has to be possible to run for all questions by the Run All in RStudio, and reproduce all of the numerics and plots in your report (up to some small randomness due to stochasticity of Monte Carlo simulations). The parts of the report that contain material that is not reproduced by the code will not be marked (i.e. the score will be 0), and the only feedback in this case will be that the results are not reproducible from the code.



c) Multiple Submissions are allowed BEFORE THE DEADLINE are allowed for both the report, and the code.
However, multiple submissions are NOT ALLOWED AFTER THE DEADLINE.
YOU WILL NOT BE ABLE TO MAKE ANY CHANGES TO YOUR SUBMISSION AFTER THE DEADLINE.
Nevertheless, if you did not submit anything before the deadline, then you can still submit your work after the deadline, but late penalties will apply. The timing of the late penalties will be determined by the time you have submitted BOTH the report, and the code (i.e. whichever was submitted later counts).

We illustrate these rules by some examples:

Alice has spent a lot of time and effort on her assignment for BDA. Unfortunately, before submission, she has accidentally introduced a typo in her code in the first question, and it did not run using Run All in RStudio. - Alice will get 0 for the questions that do not run in her code (we will try to run each code block individually), with the only feedback "Results are not reproducible from the code".

Bob has spent a lot of time and effort on his assignment for BDA. Unfortunately he forgot to submit his code. - Bob will get no personal reminder to submit his code. Bob will get 0 for the whole assignment, with the only feedback "Results are not reproducible from the code, as the code was not submitted."

Charles has spent a lot of time and effort on his assignment for BDA. He has submitted both his code and report in the correct formats. However, he did not include any explanations in the report. Charles will get 0 for the whole assignment, with the only feedback "Explanation is missing."

Denise has spent a lot of time and effort on her assignment for BDA. She has submitted her report in the correct format, but thought that she can include her code as a link in the report, and upload it online (such as Github, or Dropbox). - Denise will get 0 for the whole assignment, with the only feedback "Code was not uploaded on Learn."

3) Group work: This is an INDIVIDUAL ASSIGNMENT, like a 2 week exam for the course. Communication between students about the assignment questions is not permitted. Students who submit work that has not been done individually will be reported for Academic Misconduct, that can lead to serious consequences. Each problem will be marked by a single instructor, so we will be able to spot students who copy.

4) Piazza: During the periods of the assignments, the instructor will change Piazza to allow messaging the instructors only, i.e. students will not see each others messages and replies.

Only questions regarding clarification of the statement of the problems will be answered by the instructors. The instructors will not give you any information related to the solution of the problems, such questions will be simply answered as "This is not about the statement of the problem so we cannot answer your question."

THE INSTRUCTORS ARE NOT GOING TO DEBUG YOUR CODE, AND YOU ARE ASSESSED ON YOUR ABILITY TO RESOLVE ANY CODING OR TECHNICAL DIFFICULTIES THAT YOU ENCOUNTER ON YOUR OWN.

5) Office hours: There will be two office hours per week (Monday 14:00-15:00, and Wednesdays 15:00-16:00) during the 2 weeks for this assignment. The links are available on Learn / Course Information. I will be happy to discuss the course/workshop materials. However, I will only answer questions about the assignment that require clarifying the statement of the problems, and will not give you any information about the solutions. Students who ask for feedback on their assignment solutions during office hours will be removed from the meeting.

6) Late submissions and extensions: NO EXTENSIONS ARE ALLOWED FOR THIS ASSIGNMENT, AND THERE IS NO SUCH OPTION PROVIDED IN THE ESC SYSTEM. Students who have existing Learning Adjustments in Euclid will be allowed to have the same adjustments applied to this course as well, but they need to apply for this BEFORE THE DEADLINE on the website

https://www.ed.ac.uk/student-administration/extensions-special-circumstances

by clicking on "Access your learning adjustment". This will be approved automatically.

Students who submit their work late will have late submission penalties applied by the ESC team automatically (this means that even if you are 1 second late because of your internet connection was slow, the penalties will still apply). The penalties are 5% of the total mark deduced for every day of delay started (i.e. one minute of delay counts for 1 day). The course instructors do not have any role in setting these penalties, we will not be able to change them.

7) Please make sure to tag all pages in your submission on Gradescope, otherwise we may miss some of your work. Once your upload is complete, tagging does not counts towards your submission time (i.e. you won't get any late penalties for doing it).

```
rm(list = ls(all = TRUE))
#Do not delete this!
#It clears all variables to ensure reproducibility
```

Problem 1

In this problem, we study a dataset about car insurance. This data set is based on one-year vehicle insurance policies taken out in 2004 or 2005. In total, there are 67856 policies, of which 4624 have claims.

```
require(insuranceData)
```

```
## Loading required package: insuranceData
```

```
data(dataCar)

#You may need to set the working directory first before loading the dataset
#setwd("location of Assignment 1")
#The first 6 rows of the dataframe
print.data.frame(dataCar[1:6,])
```

```
##   veh_value  exposure clm numclaims claimcst0 veh_body veh_age gender area
## 1      1.06 0.3039014   0         0         0    HBACK       3      F    C
## 2      1.03 0.6488706   0         0         0    HBACK       2      F    A
## 3      3.26 0.5694730   0         0         0      UTE       2      F    E
## 4      4.14 0.3175907   0         0         0    STNWG       2      F    D
## 5      0.72 0.6488706   0         0         0    HBACK       4      F    C
## 6      2.01 0.8542094   0         0         0    HDTOP       3      M    C
##   agecat            X_OBSTAT_
## 1      2 01101    0    0    0
## 2      4 01101    0    0    0
## 3      2 01101    0    0    0
## 4      2 01101    0    0    0
```

```
## 5      2 01101    0    0    0
## 6      4 01101    0    0    0
```

Description of the columns.

veh_value: vehicle value in $10000s

exposure: maximum portion of the vehicle value the insurer may need to pay out in case of an incident

claimcst0: claim amount (0 if no claim)

clm: whether there was a claim during the 1 year duration

numclaims: number of claims during the 1 year duration

veh_body types: BUS = bus CONVT = convertible COUPE = coupe HBACK = hatchback HDTOP = hardtop MCARA = motorized caravan MIBUS = minibus PANVN = panel van RDSTR = roadster SEDAN = sedan STNWG = station wagon TRUCK = truck UTE = utility

gender: F- female, M - male

area: a factor with levels A,B,C,D,E, F

agecat: age category, 1 (youngest), 2, 3, 4, 5, 6

You can use either JAGS, Stan, or INLA for this question.

a)[10 marks] Fit a Bayesian logistic regression model on the dataset dataCar with

- **clm as response,**
- **a link function of your choice,**
- **using veh_value, exposure, veh_body, veh_age, gender, area, and agecat as covariates (you can use categorical covariates by converting integers to factors if appropriate).**

Center and scale the non-categorical covariates.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the linear predictor and on the response.]

Compute the posterior means of the model parameters, and discuss the results.

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation):

b)[10 marks] Fit a Bayesian Poisson regression model on numclaims as response with

- **log link function,**
- **using veh_value, exposure, veh_body, veh_age, gender, area, and agecat as covariates.**

Center and scale the non-categorical covariates.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the linear predictor and the response.]

Compute the posterior means of the model parameters, and discuss the results.

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation):

c)[10 marks] Fit a zero-inflated Bayesian Poisson regression model (https://en.wikipedia.org /wiki/Zero-inflated_model) on

- **numclaims as response,**
- **with log link function,**
- **using veh_value, exposure, veh_body, veh_age, gender, area, and agecat as covariates.**

Center and scale the non-categorical covariates.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the linear predictor and the response.]

Compute the posterior means of the model parameters, and discuss the results.

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation):

d)[10 marks] Fit a new model on numclaims in terms of the same covariates to improve on the models in part b) or part c) by considering interactions between covariates, as well as random effects. Describe your new model and justify your choices.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the linear predictor and the response.]

Compute the posterior means of the model parameters, and discuss the results.

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation):

e)[10 marks] Perform posterior predictive model checks for your models b, c, d (i.e. using replicates).

As test functions, use the number of rows in the dataset with numclaims equal 0, 1, 2, 3, and 4 (5 test functions).

Compute the RMSE values for predicting numclaims based on all 3 models.

Discuss the results.

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation):

Problem 2 - Barcelona study

In this problem, we will use a dataset from the CitieS-Health project that provides insight into the impact of air pollution on humans. It is comprised of data collected in Barcelona, Spain, and examines various environmental variables, such as air pollution levels, and their effects on mental health and wellbeing. In addition to environmental factors, this dataset also captures self-reported survey data on mental health, physical activity, diet habits, and more. From performance in a Stroop test (a type of psychological test evaluating attention capacity and processing speed) to information on total noise exposure at 55 dB - this dataset contains interesting information to understand the link between air pollution and human health.

We start by loading the dataset.

```
study<-read.csv("Barcelona.csv")
head(study)
```

```
##   Person_ID date_all year month day dayoftheweek hour sadness wellbeing energy
## 1       115    22222 2020    11   3            1   18      14         3      2
## 2       212    22247 2020    11  28            5   18       4         9      9
## 3       104    22208 2020    10  20            1   20       1         6      6
## 4       216    22247 2020    11  28            5   18       2         8      8
## 5        94    22213 2020    10  25            6   19      12         8      4
## 6       215    22258 2020    12   9            2   20       4         7      7
##   stress sleep hours_out physical_activity computer_use on_a_diet alcohol drugs
## 1      5     2         5                No          Yes       Yes      No    No
## 2      1     9         5               Yes           No        No     Yes    No
## 3      7     9        11                No          Yes       Yes      No    No
## 4      1     3         2               Yes           No       Yes     Yes    No
## 5      2     8         1                No          Yes        No      No   Yes
```

```
## 6      7     9         5               Yes           No       Yes     Yes    No
##   sick other_factors stroop_test_performance no2bcn_24h no2bcn_12h no2gps_24h
## 1   No           Yes                58.17712   33.81250  33.666667   24.32836
## 2   No            No                40.35988   15.80159  18.333333   15.48938
## 3   No           Yes                36.79430   47.52778  34.888889   48.59409
## 4  Yes            No                36.32432   15.80159  18.333333   15.64394
## 5   No            No                42.78266   12.35065   9.595238   17.03566
## 6   No           Yes                42.36540   16.91071  23.011905   22.38318
##   no2gps_12h no2bcn_12h_x30 no2bcn_24h_x30 no2gps_12h_x30 no2gps_24h_x30
## 1   22.66778      1.1222222      1.1270833      0.8109452      0.8109452
## 2   18.20557      0.6111111      0.5267196      0.5163127      0.5163127
## 3   28.62250      1.1629629      1.5842593      1.6198030      1.6198030
## 4   18.28909      0.6111111      0.5267196      0.5214648      0.5214648
## 5   15.02632      0.3198413      0.4116883      0.5678554      0.5678554
## 6   29.95232      0.7670635      0.5636905      0.7461060      0.7461060
##     pm25bcn  BCmicrog sec_noise55_day sec_noise65_day sec_greenblue_day
## 1 16.533333 1.1670614               0               0                 0
## 2  8.916667 0.2854848               0               0                 0
## 3 11.516667 1.0294803               0               0                 0
## 4  8.916667 0.2854848           37430            1426              6343
## 5 11.150000 0.4683368           12185               0                 0
## 6 10.460000 0.2532321           20596           14601                 0
##   tmean_24h tmean_12h humi_24h humi_12h pressure_24h pressure_12h precip_24h
## 1  18.05417  18.25833 82.97917 78.20833     1020.179     1020.983          0
## 2  13.89167  14.36667 86.47917 81.79167     1002.600     1001.575         37
## 3  18.98958  20.58750 76.12500 74.50000     1013.992     1012.621          0
## 4  13.89167  14.36667 86.47917 81.79167     1002.600     1001.575         37
## 5  18.57609  19.87083 51.00000 49.16667     1009.852     1007.842          0
## 6  10.19375  11.70833 47.77083 45.62500     1005.508     1006.933          0
##   maxwindspeed_24h access_greenbluespaces_300mbuff      incidence_cat age_yrs
## 1                0                             Yes       No incidence      29
## 2                4                             Yes Mobility incidence      28
## 3                0                             Yes Physical incidence      50
## 4                4                              No Mobility incidence      25
## 5                0                             Yes Physical incidence      35
## 6                0                              No       No incidence      48
##   yearbirth smoke gender     district  education microgram3
## 1      1991    No  Woman   Sant Martí University      15.72
## 2      1992    No  Woman Ciutat Vella University      37.50
## 3      1970   Yes    Man     Eixample University      41.97
## 4      1995   Yes    Man       Gràcia University      33.49
## 5      1985   Yes    Man   Sant Martí University      33.47
## 6      1972    No  Woman Ciutat Vella University      25.91
```

Descriptions of some of the covariates:

Column name	Description
Person_ID	ID of person filling out the survey (integer). Multiple rows for most persons, at different dates.
date_all	Date of the survey. (Date)
year	Year of the survey. (Integer)
month	Month of the survey. (Integer)
day	Day of the survey. (Integer)
dayoftheweek	Day of the week of the survey. (Integer)

Column name	Description
hour	Hour of the survey. (Integer)
sadness	Sadness score. (Integer)
wellbeing	Self-reported survey responses regarding wellbeing. (Integer)
energy	Self-reported survey responses regarding energy levels. (Integer)
stress	Self-reported survey responses regarding stress levels. (Integer)
sleep	Self-reported survey responses regarding sleep quality. (Integer)
hours_out	Self-reported survey responses regarding time spent outdoors. (Integer)
computer_use	Self-reported survey responses regarding computer use. (Yes/No)
on_a_diet	Self-reported survey responses regarding diet. (Yes/No)
alcohol	Self-reported survey responses regarding alcohol consumption. (Yes/No)
drugs	Self-reported survey responses regarding drug use. (Yes/No)
sick	Self-reported survey responses regarding illness. (Yes/No)
other_factors	Self-reported survey responses regarding other factors. (Yes/No)
stroop_test_performance	Performance in the Stroop test. (Float)
no2bcn_24h	Nitrogen dioxide (NO2) levels in Barcelona over 24 hours. (Float)
no2bcn_12h	Nitrogen dioxide (NO2) levels in Barcelona over 12 hours. (Float)
no2gps_24h	Nitrogen dioxide (NO2) levels in GPS locations over 24 hours. (Float)
no2gps_12h	Nitrogen dioxide (NO2) levels in GPS locations over 12 hours. (Float)
no2bcn_12h_x30	Nitrogen dioxide (NO2) levels in Barcelona over 12 hours multiplied by 30. (Float)
no2bcn_24h_x30	Nitrogen dioxide (NO2) levels in Barcelona over 24 hours multiplied by 30. (Float)
no2gps_12h_x30	Nitrogen dioxide (NO2) levels in GPS locations over 12 hours multiplied by 30. (Float)
no2gps_24h_x30	Nitrogen dioxide (NO2) levels in GPS locations over 24 hours multiplied by 30. (Float)
min_gps	Minimum GPS location. (Float)
district	District of Barcelona where the survey was conducted. (String)
education	Educational level of the participant. (String)
maxwindspeed_12h	Maximum wind speed over 12 hours. (Float)
access_greenbluespaces_300mbuff	Access to green and blue spaces within a 300m buffer. (Yes/No)
microgram3	Micrograms per cubic meter of pollutants. (Float)
age_yrs	Age of the participant in years. (Integer)
yearbirth	Year of birth of the participant. (Integer)
smoke	Self-reported survey responses regarding smoking status. (Yes/No)
gender	Gender of the participant. (Woman/Man)
hour_gps	Hour of the GPS location. (Integer)
pm25bcn	Particulate matter (PM2.5) levels in Barcelona. (Float)
BCmicrog	Black carbon (BC) levels in micrograms. (Float)
sec_noise55_day	Seconds of noise over 55 minutes in a day. (Integer)
sec_noise65_day	Seconds of noise over 65 minutes in a day. (Integer)
tmean_24h	Mean temperature over 24 hours. (Float)
tmean_12h	Mean temperature over 12 hours. (Float)
humi_24h	Humidity over 24 hours. (Float)
humi_12h	Humidity over 12 hours. (Float)
pressure_24h	Pressure over 24 hours. (Float)
pressure_12h	Pressure over 12 hours. (Float)
precip_24h	Precipitation over 24 hours. (Float)

Column name	Description
precip_12h	Precipitation over 12 hours. (Float)
precip_12h_binary	Binary value for precipitation over 12 hours. (Integer)
precip_24h_binary	Binary value for precipitation over 24 hours. (Integer)
maxwindspeed_24h	Maximum wind speed over 24 hours. (Float)

You can use either JAGS, Stan, or INLA for this question.

a)[10 marks] Fit a Bayesian linear regression model

- **on the logarithm of stroop_test_performance as response,**
- **using the following covariates: gender, on_a_diet, alcohol, drugs, sick, other_factors, educational, smoke, no2gps_24h, maxwindspeed_24h, precip_24h, sec_noise55_day, access_greenbluespaces_300mbuff, age_yrs, tmean_24h (you can use categorical covariates by converting integers to factors if appropriate).**

Center and scale the non-categorical covariates.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the response.]

Compute the posterior means of the model parameters, and interpret their meaning.

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation):

b)[10 marks] Fit a Bayesian Poisson GLM

- **for sadness as response,**
- **log link function,**
- **using the following covariates: gender, on_a_diet, alcohol, drugs, sick, other_factors, educational, smoke, no2gps_24h, maxwindspeed_24h, precip_24h, sec_noise55_day, access_greenbluespaces_300mbuff, age_yrs, tmean_24h (you can use categorical covariates by converting integers to factors if appropriate).**

Center and scale the non-categorical covariates.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the response.]

Compute the posterior means of the model parameters, and interpret their meaning.

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation):

c)[10 marks] Incorporate Person_ID as a random effects into the models a.) and b.).

Choose your own prior distributions for this random effect (do not use default priors).

Compare the posterior means of the parameter values with a) and b).

Discuss the changes that happened due to using random effects.

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation):

d)[10 marks] Do posterior predictive checks (i.e. using replicates) for the sadness score for your models with or without random effects. Explain the choice of test functions that you used.

Compute the posterior means of the response variable using the original covariates, and use this to compute the RMSE values for both models (i.e. with, or without random effects).

Discuss the results.

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation):

e)[10 marks]

Plot the posterior predictive distributions for stroop_test_performance and sadness for the random effect models in part c) for the following new person in the dataset:

Person_ID=286, gender="Woman", on_a_diet="Yes", alcohol="No", drugs="No", sick="No", other_factors="No", education="University", smoke="Yes", no2gps_24h=80, maxwindspeed_24h=10, precip_24h=50, sec_noise55_day=10000, access_greenbluespaces_300mbuff="Ye age_yrs=40, tmean_24h=25

In the case of stroop_test_performance, plot the estimated density, while for sadness, plot a histogram.

Compute the posterior predictive mean, and standard deviation.

Discuss the results.

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation):